Assessing Clustering Methods for Exploratory Spatial Data Analysis

Alan T. Murray

Australian Housing and Urban Research Institute
Department of Geographical Sciences and Planning
University of Queensland
Brisbane, Queensland 4072
Australia
(Email: alan.murray@mailbox.uq.edu.au)

Paper presented at
38th European Congress of the Regional Science Association
Vienna, Austria
August 28 – September 1, 1998

Abstract

Spatial analysis is an important area of research which continues to make major contributions to the exploratory capabilities of geographical information systems. The use and application of classic clustering methods is being pursued as an exploratory approach for the analysis of spatially referenced data. Preliminary indications are that this is both an effective and promising approach for identifying obscure or hidden attribute based patterns in spatial and non-spatial applications. However, a variety of clustering methods does exist, with different interpretations and meanings. It is essential that a better understanding of these approaches in the geographic domain be pursued in terms of their respective computational requirements and clustering implications. This paper evaluates two optimization based clustering approaches for use in the context of exploratory spatial data analysis.

Introduction

A significant amount of spatial information is being created, updated and manipulated on a daily basis. The major contributors to this spatial data explosion are geographical information systems (GIS) and remote sensing techniques which have enhanced capabilities for generating, storing and managing spatial data. Of course this is also driven by the needs of analysts, planners and policy makers who are attempting to make better and more informed decisions concerning issues such as regional growth and development, environmental sustainability, and natural resource utilization. It is one thing to have digital geographic information, but a far more challenging issue is

how this information can be understood in decision making environments. That is, what does the data indicate or suggest and what are the implications. Advanced methods for analyzing and synthesizing spatial information in a GIS environment continues to be an important area of current research.

One research focus has been on automated methods for assisting in the investigation and summarization of spatial information - exploratory spatial data analysis (ESDA). Clustering techniques have emerged as a potential approach for analyzing complex spatial data in order to determine whether or not inherent geographically based relationships exist. One example is the application of clustering to identify and detect potential cancer or disease patterns in populations based on the analysis of a set or subset of spatial attributes. Another application is the analysis of criminal offenses where trends in occurrence as well as potential shortcomings in policing practices are of interest. The use of clustering in the spatial domain is currently based on notions of pattern spotting and data mining. This is a natural progression of the use of classic statistical approaches for hypothesis testing.

There are a number of alternative optimization based modeling approaches for identifying clusters in spatial data. A recent review of these approaches may be found in Murray and Estivill-Castro (1998). Two approaches will be investigated in this paper. The first is based on the spatial modeling work of Cooper (1963), which may be considered a geographically sensitive variant of the k-means approach (MacQueen 1967) found in most, if not all, statistical software packages. In this paper this approach is referred to as the center points clustering problem. The second approach is based on the spatial analysis research of Hakimi (1965), which is actually a spatial extension of the grouping work of Vinod (1969). In this paper this approach is denoted the median clustering problem. The focus of this paper is on the need to better understand the similarities and differences between these two alternative clustering approaches, primarily in terms of the spatial ramifications of identified clusters.

This paper begins by detailing the center points and median clustering models. Spatial groups identified by these approaches are then investigated. One focus of comparison is on functional differences between the center points and median model objective

measures. The second focus of evaluation contrasts produced cluster groupings in order to develop an understanding of possible spatial variation. Finally, a discussion and conclusions are given.

Clustering Models

There are three general optimization based models which may be applied to identify clusters in spatial information (Murray and Estivill-Castro 1998). The differences between these approaches are in how clusters are defined and evaluated. Two of these approaches will be evaluated in this paper: the center points approach (i.e. Cooper 1963) and the median approach (i.e. Hakimi 1965).

Clustering using center points enables groupings of observation sites to be identified based upon the use of artificial points in space. These points, or center points, serve as a means for creating spatial clusters. The following notation will be used in the specification of this clustering approach:

i = index of observation sites (total number = n);

k = index of center points (total number = p);

d_{ik} = spatial difference relating observation i and center point k;

$$y_{ik} = \begin{cases} 1 & \text{if observation site } i \text{ is in cluster } k \\ 0 & \text{otherwise.} \end{cases}$$

It is worth pointing out that there are a number of ways to specify the spatial difference measure. Throughout this paper, the spatial difference is defined to be the Euclidean distance measure.

Center Points Clustering Problem (CPCP)

Minimize

$$Z = \sum_i \sum_k a_i d_{ik} y_{ik} \qquad (1)$$

Subject to:

$$\sum_k y_{ik} = 1 \qquad \forall i \tag{2}$$

$$y_{ik} = (0,1) \qquad \forall i,k \tag{3}$$

The objective (1) of the CPCP is to minimize the total difference in the assignment of observation sites to cluster center points. Unfortunately, determining the location of the center points is also a significant component of the problem. Thus, the objective is non-linear as it is based upon the use of center point location decision variables and is particularly difficult to solve (see Rosing 1991). Constraint (2) ensures that observation sites are assigned to a cluster. Constraint (3) imposes integer restrictions on decision variables.

The CPCP is well known in the statistics literature as the k-means clustering approach (see MacQueen 1967), where a distance squared difference measure, d_{ik}^2, is utilized in objective (1). Given this, the center points in the k-means approach correspond to the cluster centroids. It has been shown in Murray and Estivill-Castro (1998), among others, that the use of the distance squared measure, and hence the centroid, is problematic and should not be utilized for either spatial or aspatial clustering applications.

An alternative to the use of center points is to use the spatial observations themselves as a means for identifying spatial clusters. The use of observations corresponds to a median in the location literature as discussed in Murray and Estivill-Castro (1998), which distinguishes this as a median clustering approach. The following notation will assist in the specification of this alternative clustering model:

i = index of observation sites (total number = n);

j = index of potential medians (same as i);

d_{ij} = spatial difference between observation i and potential median j;

p = number of cluster medians to be selected;

$$x_j = \begin{cases} 1 & \text{if cluster median } j \text{ is selected} \\ 0 & \text{otherwise.} \end{cases}$$

$$z_{ij} = \begin{cases} 1 & \text{if observation } i \text{ is assigned to cluster median } j \\ 0 & \text{otherwise.} \end{cases}$$

Median Clustering Problem (MCP)

Minimize
$$Z = \sum_i \sum_j a_i d_{ij} z_{ij} \tag{4}$$

Subject to:

$$\sum_j z_{ij} = 1 \qquad \forall i \tag{5}$$

$$z_{ij} \leq x \qquad \forall i, j \tag{6}$$

$$\sum_j x_j = p \tag{7}$$

$$z_{ij} = (0,1) \qquad \forall i, j \tag{8}$$
$$x_j = (0,1) \qquad \forall j$$

The objective (4) of the MCP is to minimize the total weighted grouping of observation sites to selected medians. Constraint (5) ensures that observation sites are assigned to a median. Constraint (6) requires a median to be selected before it may serve as a representative location for grouping observation sites. Constraint (7) specifies that *p* clusters be identified. Constraint (8) imposes integer restrictions on decision variables.

The two clustering models are very much related in that they both use locations in space to create clusterings. Unfortunately, the CPCP uses points which are a function

of individual cluster membership. Alternatively, a median location is predefined as it is a member of the observation sites. This distinction is important and does have implications for the computational difficulty one can expect in solving either the CPCP or MCP using exact or heuristic solution techniques. Beyond this, there is currently no understanding of the similarities or differences between these two approaches in terms of the clusterings produced – functionally or spatially. The major focus of this paper is to develop a better understanding of such relationships and properties within the context of ESDA.

Obtaining Cluster Solutions

Both the CPCP and the MCP are difficult and challenging problems to solve either by exact or heuristic methods. Exact solution techniques for the CPCP are limited to relatively small problem applications (see Rosing 1992), so heuristic approaches are essential. The most widely acknowledged and applied heuristic for the CPCP is the alternating approach developed by Cooper (1964):

(i) Generate p clusters.

(ii) Identify a representative center point for each cluster.

(iii) Assign observation sites to their closest center point.

(iv) If the cluster groupings have changed, then return to (ii). Otherwise, a local optima has been reached and the heuristic terminates.

Others heuristics for the CPCP have been developed and applied (Cooper 1967; Love and Juel 1982; Houck et al. 1996), but it is not clear whether they identify solutions of higher quality using spatial information than the original alternating heuristic. Given this, the alternating heuristic for the CPCP was utilized for obtaining application solutions. The best solution identified is reported for the CPCP from 10,000 randomly generated initial clusterings to which the alternating heuristic was applied. For the MCP, both exact and heuristic solution techniques have been developed and applied to medium and large problem instances (see Murray and Church 1996). Solutions reported here for the MCP have been identified using Lagrangian relaxation with branch and bound and are optimal to within 0.001%. Details on this approach for

solving the MCP and extensions of the MCP may be found in Murray and Gerrard (1997).

Three spatial applications have been solved on a Pentium II/300 personal computer for a range of *p* cluster values. The first application contains 33 observation sites from Austin, Texas and represents emergency service calls in this region (Daskin 1982). The second application contains 55 observation sites in Washington D.C. (Swain 1971). The final application contains 152 observation sites from the Busoga, Uganda region, representing coffee buying centers (Migereko 1983).

Reported in Tables 1-3 are objective function values for the CPCP and the MCP using the three detailed spatial applications for a range of cluster values. For each value of *p*, the tables report the best cluster solution found by the two approaches, identified by the shadowed boxes, as well as the functional value of this solution evaluated using the other clustering model. For example, the best clustering found for the CPCP in Table 2 for p=5 has an objective (1) functional value of 2927.46, as indicated in the shaded box, and evaluating this clustering using the MCP results in an objective (4) functional value of 2945.70. The optimal clustering found for the MCP is given below this in the shaded box, having an objective (4) functional value of 2944.20 and evaluating this clustering using the CPCP indicates an objective (1) functional value of 2928.32. Thus, each value of *p* indicates the best solution found for each clustering model (in the shaded box) as well as an evaluation of the clustering solution identified using the other model. Given this reporting scheme, what would be expected is that the shaded box always indicates a superior value (lower) for the corresponding column heading and value of *p* than the evaluated clustering identified by the other approach.

The clusters identified in Table 1 for each value of *p* are identical and result in the same objective function measures when evaluated as either a CPCP or a MCP. Thus, the CPCP solution for p=4 in the shaded box of 10,741.19 is repeated directly below as the MCP identified the same clusters. This is not the case, however, for any of the cluster solutions reported in Tables 2 or 3. Solution times for the problems summarized in Table 1 were less than 0.02 seconds per solution using the alternating

Table 1. Clustering solutions using the 33 observation site data.

	CPCP	MCP
p=3	12,434.76	12,434.67
	12,434.76	12,434.67
p=4	10,741.19	10,741.28
	10,741.19	10,741.28
p=5	9312.08	9316.69
	9312.08	9316.69
p=6	8180.67	8196.81
	8180.67	8196.81

Table 2. Clustering solutions using the 55 observation site data.

	CPCP	MCP
p=5	2927.46	2945.70
	2928.32	2944.20
p=6	2635.87	2651.33
	2636.94	2649.55
p=7	2417.06	2422.30
	2417.93	2420.79
p=8	2203.55	2230.08
	2211.78	2217.85
p=9	2061.11	2091.49
	2065.12	2071.19
p=10	1922.85	1936.66
	1919.67	1927.45

Table 3. Clustering solutions using the 152 observation site data.

	CPCP	MCP
p=5	466,719.80	470,516.13
	466,764.80	470,093.78
p=6	423,797.60	427,768.25
	424,304.30	425,847.28
p=7	385,720.80	390,416.03
	386,876.50	388,914.28
p=8	358,228.00	361,591.66
	358,417.80	361,348.88
p=9	334,670.00	339,607.31
	335,116.20	337,015.03
p=10	315,031.84	318,836.72
	315,728.80	318,666.66
p=11	297,839.25	300,477.13
	298,340.64	300,422.96
p=12	282,215.28	285,129.28
	282,642.30	284,857.78
P=13	265,034.69	267,889.06
	265,371.15	266,991.76
p=14	250,718.97	252,753.78
	250,728.52	252,392.95
p=15	242,202.94	245,388.70
	238,852.89	240,438.35

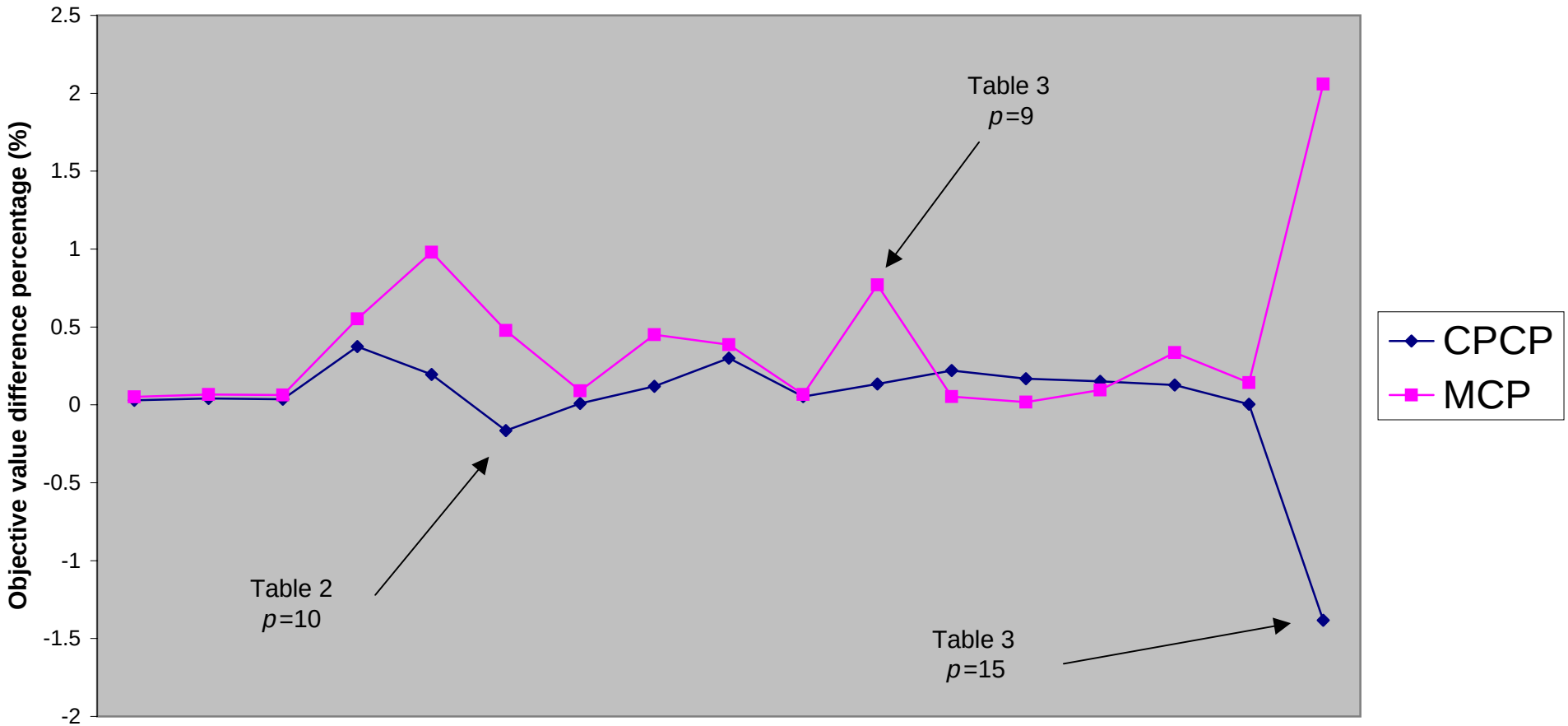
heuristic for the CPCP and less than 0.03 seconds using the Lagrangian relaxation approach for the MCP for each value of *p*.

Table 2 differs substantially from Table 1 in that the two approaches do not identify the same clusterings for any of the *p* values. Given this, it is interesting that there is a significant amount of agreement between the CPCP and MCP in terms of each approach identifying solutions which functionally rate well when evaluated using the other model. As an example, for p=7 in Table 2 the CPCP identifies a clustering which gives a functional value of 2417.06 and the optimal MCP grouping evaluated as a CPCP results in a functional value of 2417.93. Alternatively, the MCP identifies a clustering which gives a functional value of 2420.79 and the best CPCP clustering evaluated as a MCP results in a functional value of 2422.30. It is important to note that for p=10 in Table 2, the MCP actually identified a clustering which was superior to the clustering identified by the alternating heuristic for the CPCP. Specifically, the alternating heuristic identified a clustering with a CPCP functional value of 1922.85, whereas the optimal MCP grouping evaluated as a CPCP has a functional value of 1919.67. Obviously the alternating heuristic did not find a globally optimal solution for the CPCP in this instance. Solution times for the problems summarized in Table 2 were less than 0.04 seconds per solution using the alternating heuristic for the CPCP and less than 0.95 seconds using the Lagrangian relaxation approach for the MCP for each value of *p*.

The results found in Table 3 are quite similar to those reported in Table 2. Significant agreement between the two approaches would appear to exist. Further, there is also an instance in Table 3 where the MCP identifies a functionally better solution for the CPCP than does the alternating heuristic (p=15). The alternating heuristic obviously was not able to find a globally optimal solution in this case. Solution times for the problems summarized in Table 3 were less than 0.17 seconds per solution using the alternating heuristic for the CPCP and as high as 34.10 seconds using the Lagrangian relaxation approach for the MCP for each value of *p*.

In order to summarize the findings given in Tables 2 and 3, Figure 1 shows the objective function difference percentages between the identified and evaluated clusters. Specifically, for the CPCP this is the difference between the best clustering

Figure 1. Functional value differences for the results given in Tables 2 & 3.



found and the evaluated MCP grouping. For the MCP, this is the difference between the optimal MCP grouping and the evaluated CPCP grouping. The first 6 plotted differences correspond to p=5-10 in Table 2 and the last 11 plotted differences are associated with p=5-15 in Table 3. Examining the highlighted p=9 entry from Table 3 in Figure 1 shows that the best CPCP grouping evaluated as an MCP deviates from the optimal MCP solution by 0.77%. Alternatively, the optimal MCP grouping evaluated as a CPCP is 0.13% higher than the best CPCP solution. The only major CPCP deviation in Figure 1 is attributed to the sub-optimal clustering found for the CPCP in Table 3 for p=15. Given this, the MCP may be considered a very good model for identifying clusters which will be high quality or optimal CPCP solutions. The CPCP groupings are also high quality MCP solutions, but perhaps not to the extent previously discussed.

Cluster Evaluation

As discussed previously, the functional evaluation and comparison of the MCP and the CPCP groupings is very important, but the spatial extent of any differences is certainly of particular interest in this paper. Examining spatial differences in clustering solutions is a challenging task as they are difficult to represent, interpret and summarize. The reason for this is that a spatial model, in general, may have a large number of solutions (clusterings in this case) which are functionally similar, but are very different from each other spatially. Thus, if spatial patterns are alike then it is safe to draw conclusions from such an occurrence. However, if spatial patterns are disparate then this may be an artifact of the solution space, which is partially defined by the model objective function (equation 1 for the CPCP and equation 4 for the MCP). Given this, the strongest case for establishing that two approaches are producing similar clusterings is where the groupings evaluate favorably in terms of the objective function measure and have a very comparable spatial pattern.

The previous section has established that the groupings produced by the MCP and the CPCP for the three reported spatial applications demonstrate functional similarity. This point is supported by the results summarized in Figure 1, where the objective function percentage deviations for the identified clusters remain consistently low. In fact, the worst comparative evaluation deviated by approximately 2%. A number of

these clusters will now be compared using a basic and spatially explicit evaluation approach.

Figure 2 displays the five clusters (p=5) identified for the Swain application reported in Table 2. The solid lines represent MCP groupings and the dashed lines represent CPCP groupings. What is shown in Figure 2 is that the groupings identified by the two clustering models vary by only one observation site. Specifically, observation site 49 has a different group membership when comparing the MCP groupings to the CPCP groupings. Barring this, the cluster boundaries coincide for the two approaches. It is fairly obvious that the clusters are similar in this case.

The nine groups (p=9) found for the MCP and the CPCP using the Swain application are presented in Figure 3. The clusters displayed in Figure 3 are more varied than those shown in Figure 2. For example, observation sites 2, 9, 25 and 39 largely represent the major changes in cluster membership. Beyond this, the CPCP, in contrast to the MCP, has combined observation sites 2, 4 and 42 into one group and has split the MCP grouping of observation sites 1, 5, 11, 13, 40, 43, 44, 46, 47, 52, 53, 55 into two groups. Although the clusters are not as similar as those found in Figure 2, there is certainly a significant amount of commonality between the identified clusters. Figure 3 is in fact the most spatially disparate clustering found for the two approaches.

The six groups (p=6) identified for the MCP and the CPCP for the Uganda application are presented in Figure 4. Only five of the 152 observation sites change cluster memberships in Figure 4. Specifically, there is a change for observation sites 18, 24, 45, 77 and 83. Otherwise, there is again almost complete boundary overlap between the two sets of produced clusters. This is another instance where cluster similarity is clearly present.

The final comparison is displayed in Figure 5 for the thirteen groups (p=13) associated with the Uganda application. In this case, only three observation sites change cluster membership between the two models. Specifically, observation sites 18, 41 and 150 are members of different groups in the CPCP as compared to the

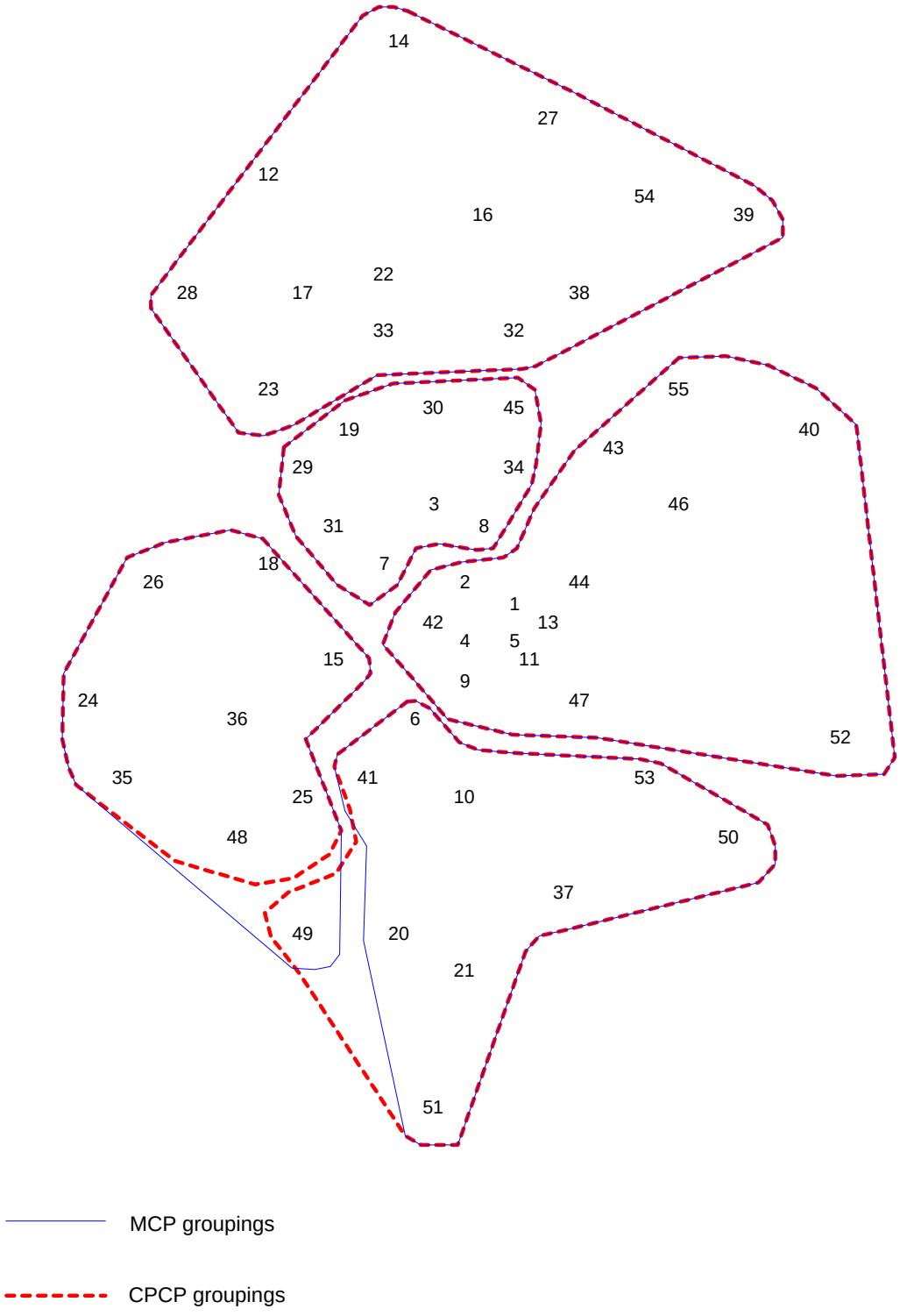


Figure 2. Five cluster grouping for the Swain application.

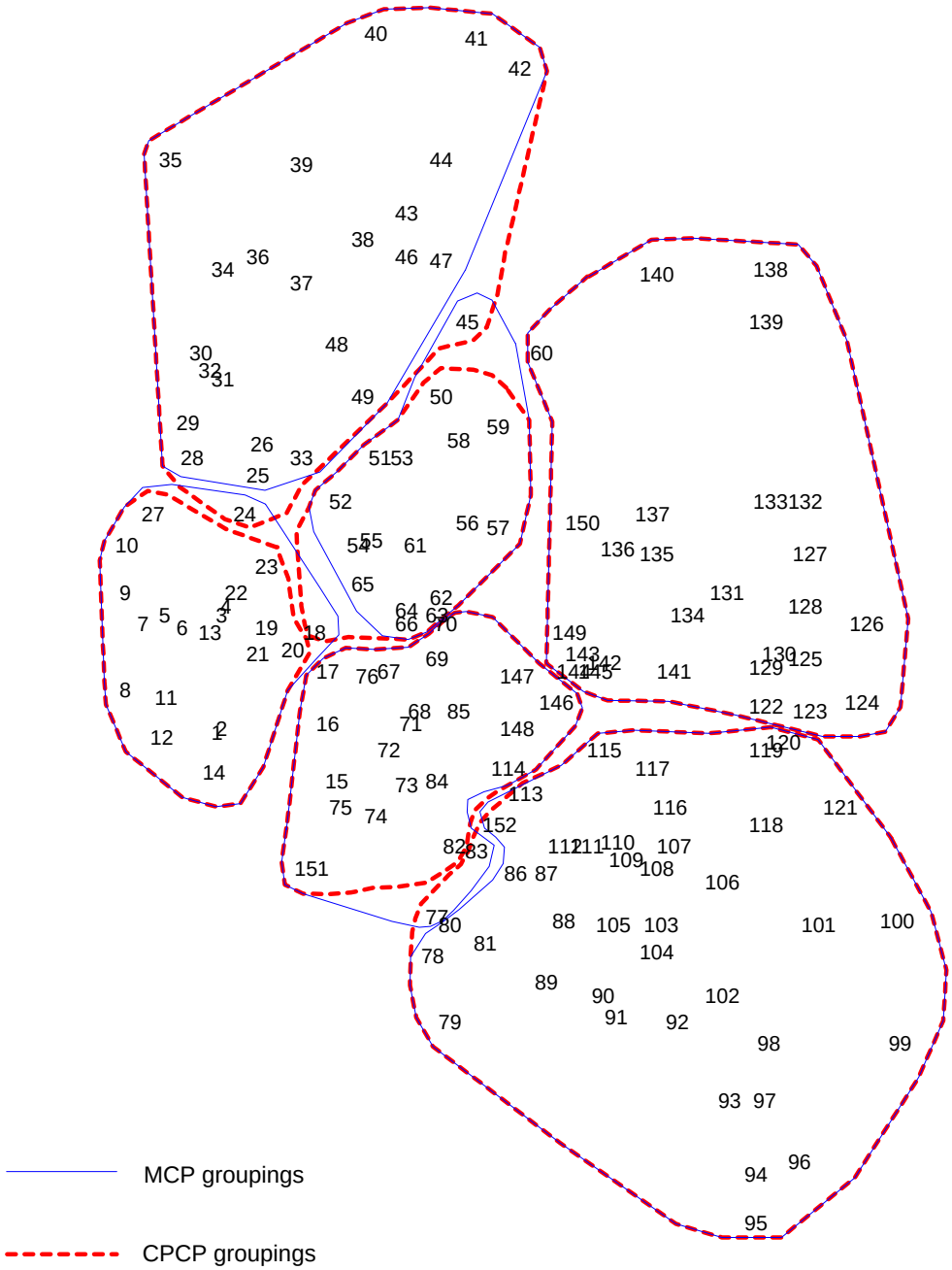


Figure 4. Six cluster grouping for the Uganda application.

MCP. Here again, there is significant visible agreement between the produced clusters.

Discussion and Conclusions

There is a strong case to support the contention that there is substantial similarity in produced clusters using the CPCP and the MCP approaches. This being the case, it is then a question of preference and computational requirements which dictate the appropriateness of these two clustering approaches for a particular spatial application. That is, all else being relatively equal, the selection of a particular clustering approach for ESDA would be a matter of individual or institutional choice. Before going further, it is worth discussing some important points regarding various aspects of the results reported thus far. This may make the differences between these alternative approaches more clear and apparent.

It is always important to have an understanding of heuristic performance when they are utilized for obtaining solutions. The major question being whether or not the solutions identified are of high quality (optimal or near optimal). Previous studies of the performance of the alternating heuristic for the CPCP indicate that it does quite well. In general, the results observed in this paper suggest that the performance of the alternating heuristic is good. However, based upon the findings presented in Tables 2 and 3, the alternating heuristic for the CPCP is not always capable of findings solutions which may be identified using an indirect approach, namely the MCP. That is, in Table 2 for p=10 and in Table 3 for p=15 the MCP identified groupings which when evaluated using the CPCP were superior to the best solutions obtained using 10,000 runs of the alternating heuristic. Another point is that the solutions reported for the CPCP in Table 2 for p=9 and in Table 3 for p=14 are not optimal, but they are better than the identified MCP groupings. Based upon these findings, the alternating heuristic for the CPCP appears to have difficulty identifying globally optimal solutions as the number of clusters increases, at least in the context of ESDA. It is interesting to note that results based upon the use of randomly generated data do not suggest this for the alternating heuristic, so this is an important observation. Heuristic development for the CPCP remains an open area of research from an optimization perspective whether the application is in the area of clustering based ESDA or the original spatial modeling oriented analysis.

Related to the issue of heuristic development is the computational performance of utilized techniques for solving either the CPCP or the MCP. The solution time comparisons noted previously suggest that at present the Lagrangian solution approach for the MCP is relatively efficient in comparison to the alternating heuristic for the CPCP. Recall that the most difficult problem applications reported in Table 3 were solved optimally in a maximum of 34.10 seconds for the MCP, whereas the 10,000 runs of the alternating heuristic for the CPCP required approximately 1700 seconds to solve. Further, as noted previously, the CPCP solutions were not necessarily optimal. At least four out of the 21 problem instances solved are known to be sub-optimal for the CPCP.

Analyzing the CPCP and the MCP in terms of functional and spatial differences provided strong evidence for the two approaches producing similar clusterings. The functional evaluation and comparison summarized in Figure 1 highlights the fact that the two clustering approaches consistently identified groupings of high quality according to both modeling approaches. The spatial comparison of the MCP and the CPCP demonstrated that the produced clusters were not particularly different, even in the worst instance. One question that may be worth pursuing further is whether there exists a technique which may be used to assess the significance of cluster similarity between alternative modeling approaches? This is certainly a valuable component for carrying out such a comparative analysis. However, the fact that clustering approaches, in general, partition space, this enables identified clusters to be contrasted visually as was done in Figures 2-5.

This paper has compared two clustering approaches which may be utilized for exploratory spatial data analysis (ESDA). The first approach was the center points clustering problem (CPCP) and the second approach was the median clustering problem (MCP). A significant amount of similarity was demonstrated between the two approaches in terms of their functional performance as well as the spatial comparability of identified clusters. The use of the MCP for ESDA would be recommended based upon the lack of any notable difference between groupings identified by either the MCP or the CPCP. In fact, the MCP appears to consistently identify optimal or near optimal CPCP groupings. Further, the fact that solving the

MCP required less computational effort than the CPCP and can be effectively solved for optimal solutions when significantly larger spatial applications are pursued supports this position as well. In a geographical information system (GIS) environment these are certainly important considerations.

Acknowledgments

This research was supported in part by a grant from the Australian Research Council.

References

Cooper, L. (1963). "Location-allocation problems." *Operations Research* **11**, 331-343.

Cooper, L. (1964). "Heuristic methods for location-allocation problems." *SIAM Review* **6**, 37-53.

Cooper, L. (1967). "Solutions of generalized locational equilibrium models." *Journal of Regional Science* **7**, 1-18.

Daskin, M. (1982). "Application of an expected covering model to emergency medical service system design." *Decision Sciences* **13**, 416-439.

Hakimi, S.L. (1965). "Optimum distribution of switching centers in a communication network and some related graph theoretic problems." *Operations Research* **13**, 462-475.

Houck, C., Joines, J. and Kay, M. (1996). "Comparison of genetic algorithms, random restart and two-opt switching for solving large location-allocation problems." *Computers and Operations Research* **23**, 587-596.

Love, R. and Juel, H. (1982). "Properties and solution methods for large location-allocation problems." *Journal of the Operational Research Society* **33**, 443-452.

MacQueen, J. (1967). "Some methods for classification and analysis of multivariate observations." In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probably*, vol. I, edited by L. Le Cam and J. Neyman, 281-297 (Berkeley: University of California Press).

Migereko, D. (1983). *An analysis of the coffee cooperative marketing system in Busoga, Uganda: transportation and facilities location*. Masters thesis. University of California, Santa Barbara.

Murray, A. and Church, R. (1996). "Applying simulated annealing to location planning models." *Journal of Heuristics* **2**, 49-71.

Murray, A. and Gerrard, R. (1997). “Capacitated service and regional constraints in location-allocation modeling.” *Location Science* **5**, 103-118.

Murray, A. and Estivill-Castro, V. (1998). “Cluster discovery techniques for exploratory spatial data analysis.” *International Journal of Geographical Information Science* **12**.

Rosing, K. (1991). “Towards the solution of the (generalised) multi-Weber problem.” *Environment and Planning B* **18**, 347-360.

Rosing, K. (1992). “An optimal method for solving the (generalized) multi-Weber problem.” *European Journal of Operational Research* **58**, 414-426.

Swain, R. (1971). *A decomposition algorithm for a class of facility location problems.* Ph.D. dissertation. Cornell University. Ithaca, NY.

Vinod, H. (1969). “Integer programming and the theory of grouping.” *Journal of the American Statistical Association* **64**, 506-517.